Principal Component Analysis (PCA)

PCA is a popular dimensionality reduction technique for analyzing large datasets containing a high number of dimensions/features per observation, increasing the interpretability of data while preserving the maximum amount of information.

Given an unsupervised dataset, the steps to be performed are:

- compute the covariance matrix S;
- compute the highest eigenvalues and the corresponding eigenvectors;
- project all the datapoints in this dimensions;

What happens when you have a small dataset?

Small number of D-dimensional samples, with $D \gg N$

e.g. a small set of high resolution images

In such cases it may be inefficient to find eigenvalues of S.

Instead of considering the covariance matrix $S = X^T X$ we consider the matrix XX^T. The two have the same eigenvalues but $X^T X$ is $D \times D$ and XX^T is $N \times N$.

$$\text{Solution: } \frac{1}{N} XX^T \underbrace{(X u_i)}_{v_i} = \lambda_i \underbrace{(X u_i)}_{v_i}$$

$$\frac{1}{N} XX^T v_i = \lambda_i v_i$$

Once we compute the eigenvalues we can compute the eigenvectors.
We have very efficient ways of computing PCA even with high dimensional data.

$$u_i = \frac{1}{\sqrt{N \lambda_i}} X^T v_i$$

Probabilistic PCA

Assume we have the input, $x \in \mathbb{R}^D$
Define another set of variables, $z \in \mathbb{R}^M$ (reduced space). We don't know these variables

Assume the conditional probability distribution $P(x|z)$ is given by a linear-gaussian model, a gaussian over x centered in a linear combination of z:

$$P(x|z) = \mathcal{N}(x \mid Wz + \mu, \sigma^2 I)$$

If we have the latent variable z and the parameters of the model μ, σ then we can generate a distribution of x.

Probabilistic PCA is just a method that, given the dataset, estimates the parameters W, μ, σ. We use the maximum likelihood technique:

$$\underset{W,\mu,\sigma}{\operatorname{argmax}} \; P(X \mid W, \mu, \sigma^2) = \sum_{n=1}^{N} \ln P(x_n \mid W, \mu, \sigma^2)$$

Setting the derivative to 0 we have a closed form solution that depends on the eigenvalues and eigenvectors of S. The proof is not trivial.

Once we have W, μ, σ we can generate new samples of x (big space) given z (latent space - reduced space).